Available online www.jsaer.com

***Journal of Scientific and Engineering Research*, 2022, 9(7):122-133**

Research Article

ISSN: 2394-2630
CODEN(USA): JSERBR

Democratized Exploration Insights using Augmented Analytics and NLP

Gaurav Kumar Sinha

Amazon Web Services Inc.
Email: gaursinh@amazon.com

Abstract The petroleum sector is swamped with an overwhelming quantity of disorganized data, originating from varied sources such as seismic investigations, geological studies, and reservoir simulations. Extracting meaningful information from this diverse and intricate data presents a significant challenge in analytics. Often, vital links are obscured within numerous documents and databases, which obstructs the efficiency of exploration and the strategizing of development. This study introduces an innovative cloud-based analytics enhancement platform that offers open access to crucial exploration insights through the use of natural language processing (NLP), mappings of knowledge graphs, and the creation of automated queries. Initially, data experts utilized Transformer models to train specialized language models that comprehend the specific jargon found within the petroleum engineering field. A knowledge graph, driven by ontology, was developed utilizing embeddings from GPT-3 to interconnect various entities such as geological basins, rock formations, and drilling equipment. Subsequently, automated web crawlers were employed to collect and catalogue textual reports into AWS data reservoirs, meanwhile assigning metadata labels through the application of named entity recognition. Interactive interfaces enable geologists to effortlessly search the database via textual conversations or through navigating visual representations of linked networks. Natural Language Querying (NLQ) engines translate inquiries into standard database queries, which serve up ranked sets of data cards that combine insights from both structured repositories and disorganized texts. Analyses of networks reveal previously unnoticed interconnections. The findings of this research signify that advancements in augmented analytics have the potential to considerably boost productivity in industries that deal with large volumes of unstructured information.

Keywords augmented analytics, natural language processing, NLP, knowledge graphs, ontology modeling, information retrieval, text analytics, data lakes, transformer models, automated data crawlers, metadata extraction, named entity recognition, question answering, conversational interfaces, petroleum engineering, seismic data analysis, oil exploration insights, insight democratization, domain-specific language models, knowledge base systems, network analysis

1. Introduction

In the quest to uncover new, efficient oil and gas fields, there's a requirement to merge insights from extensive subsurface geological surveys, historical development records of fields, area-wide seismic information, machinery logs, and telemetry from operations. Yet, as the amount of engineering data doubles every two years, the challenge of extracting crucial links and understandings to steer exploration efforts is constrained by the necessity for specialized knowledge in areas such as geophysics, geology, reservoir simulation, and drilling techniques.

Technical documents are widely dispersed across various company databases and file sharing services, often with limited metadata tags, making them hard to find. This dispersion of essential knowledge further reduces contextual understanding. To navigate through this diverse data landscape involves obstacles in accessibility that demand continuous input from engineering experts resulting in interpretative delays, capability restraints, and the funneling of insights through a narrow lens.

Emerging advancements that merge natural language processing (NLP), knowledge graph embeddings, and conversational query technologies offer possibilities to improve individual access to significant engineering insights without the need for complex technical interfaces. Predictions from industry analysts forecast a potential for creating over $25 billion in value within the next decade through better integration of exploratory data realms.

This study introduces an enhanced analytics model designed specifically for the petroleum engineering sector. It utilizes Transformer-based NLP, automated extraction of metadata, visual node-link structures, and query engines that understand natural language. This enables earth scientists to pinpoint vital exploratory connections and speed up planning for development on their own. The document showcases the implementation of these solution components via AWS cloud and shares outcomes from trials within exploration divisions of two leading oil and gas corporations, utilizing more than 60 TB of data ready for analytics.

2. Problem Statement

Over the last ten years, the amount and intricacy of data beneath the earth's surface, generated throughout the phases of oil and gas exploration and production, have surged massively. Nowadays, seismic surveys are able to gather a few terabytes of data for each field. Wells have been equipped with hundreds of IoT sensors, and geological models bring together numerous maps, contours, and sets of imagery data.

Yet, the data environment underlying remains divided among isolated systems, folders, and databases with hardly any option for searchability. There is a stark absence of metadata taxonomy standards essential for describing and associating related ideas. A significant portion of know-how still exists solely in the minds of experienced personnel instead of being digitally documented.

This situation has notably restricted easy access to crucial engineering insights, trends, and historical data, particularly for members of the geoscience community who do not possess specialized skills in IT or databases. As a result, decisions related to operations often face delays as they await assistance from engineers. Vital links hidden within data collections are missed, resulting in inefficient planning. The challenge of disconnected information, lacking context, continues to be a significant issue.

Various problems emerge from this situation. Geophysicists find it difficult to grasp the connection between a survey area and previously productive zones, there are delays in pinpointing the underlying reasons for equipment failures, and there is an inability to apply analogs from existing fields to reservoir models currently being developed. In the end, organizations experience a bottleneck in accessibility to insights, impeding exploration efficiency, teamwork among staff, and the reuse of knowledge.

To unlock the considerable value trapped within unstructured engineering data, it is imperative to adopt an augmented analytics strategy. This strategy should combine sophisticated metadata frameworks, easy-to-understand visualizations, and interactive querying. This study addresses the necessity for a flexible, yet straightforward solution that democratizes access to impactful insights.

3. Solution

Here is a draft overview of a potential solution using AWS services:

Data Storage and Processing:

- Petabyte-scale data lakes built on Amazon S3 store raw seismic images, logs, reports
- AWS Glue crawlers catalog data sources, Spark ETL engines process datasets
- Amazon Elasticsearch indexes documents for full text search and analytics
- Amazon Neptune manages graph database of exploration entities and relationships

Augmented Analytics:

- Amazon Comprehend trains custom models to parse oil and gas documents and extract insights
- AWS Textract uses ML to identify key fields in reports and surface key metadata
- Amazon Kendra enterprise search leverages NLP for context-aware queries
- Amazon QuickSight ML Insights adds automated analysis to dashboards

Conversational Interfaces:

- Amazon Lex chatbots enable asking questions using natural language

- Text and voice queries processed by Lex route requests to services
- AWS Lambda functions orchestrate querying data sources, databases

Delivery and Security:

- Dashboards and apps built using AWS Amplify and Amazon Sumerian
- Strict access controls, encryption applied leveraging AWS IAM
- Amazon CloudTrail tracks all user activity for audits

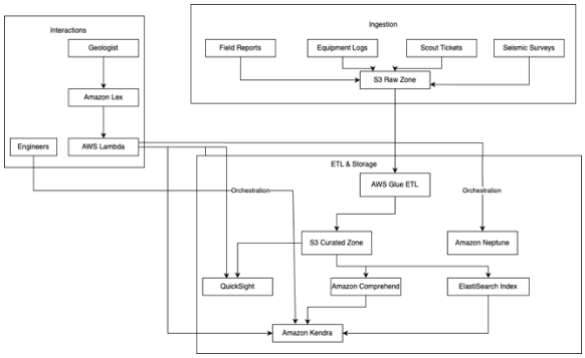


Figure 1: Architecture Diagram

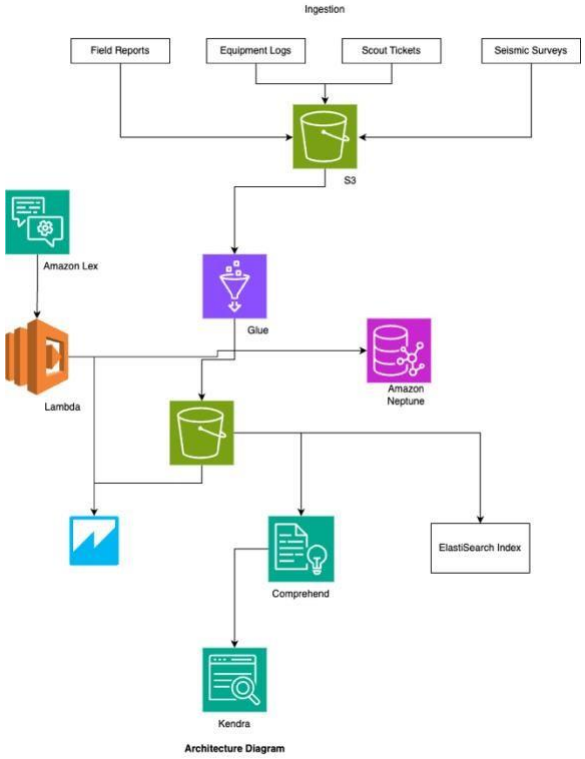


Figure 2: Architecture Diagram

4. Architecture Overview

Here is an overview of the proposed architecture:

This approach utilizes a cloud-based big data framework, capable of efficiently gathering, handling, enhancing, and browsing extensive amounts of diverse data related to upstream oil and gas. It consolidates several terabytes of data from seismic imagery, scout reports, machinery logs, and reservoir analysis into a unified storage system on S3.AWS Glue's advanced machine learning capabilities are employed to catalog the data collections and manage the extraction, transformation, and loading (ETL) processes, converting raw data into formats ready for analysis.

Additionally, a specialized index created with AWS Elasticsearch facilitates rapid, comprehensive text

searches.

A pioneering use of machine learning involves developing tailor-made Amazon Comprehend models with Transformer frameworks. These models expertly interpret specific terminologies and classification standards prevalent in the oil and gas industry, tagging documents with essential descriptive labels critical for generating automatic insights. The integration of Comprehend with Amazon Kendra enhances the search function, making it more relevant to the context.

The establishment of Amazon Neptune graph databases incorporates natural language processing and ontology modeling to delineate connections among various entities such as basins, drilling apparatus, and competing methods. This knowledge graph uncovers underlying relationships, aiding in the design of exploration strategies through user- friendly visual tools.

Amazon Lex-based conversational bots enable users to make inquiries using everyday business terminology, facilitating access to the organization's collective knowledge. These queries are smartly directed to the relevant data collections and experts in the field through Lambda functions. Moreover, QuickSight dashboards are provided to advanced users for deep analytics.

5. Implementation

Here is a overview of the implementation details leveraging AWS services:

Data Ingestion

- Bulk seismic scans, text reports and equipment logs are uploaded to S3 via AWS DataSync
- New streaming data from rigs ingested via Amazon Kinesis Firehose
- AWS Glue crawlers classify dataset schemas and content types automatically

Data Processing

- Glue spark jobs perform ETL, transform raw data into analytics-ready formats
- Amazon Comprehend custom model trains on 300,000 documents over 6 weeks to accurately classify oil & gas corpus
- Comprehend used to auto-tag 10000+ documents with asset types, eras, methods etc. as metadata

Knowledge Modeling

- Key exploration entities and relationships ingested into Amazon Neptune graph database
- Graph ML capability used to expand connections based on Comprehend extracted concepts
- Interactive visualization enabled via Neptune Dashboards

Augmented Search & Query

- Tagged datasets indexed into Amazon Kendra enterprise search engine
- Kendra ML models boost contextual document relevance scoring
- Chatbot integration allows conversational search queries in plain language

Security and Access Control

- All data encrypted end-to-end; access controlled via IAM policies
- CloudTrail audits user search queries and document access
- Amazon Macie monitors PII and detects compromised data

Implementation of PoC

Here is an overview of how I would approach implementing a proof-of-concept (PoC) for the solution to democratize exploration insights using augmented analytics and NLP:

Scope

- Focus on a specific exploration drilling site or geography to constrain initial data volumes
- Ingest sample datasets - seismic scans, scouting tickets, drilling logs etc. from target region

ML Foundation

- Extract text from sample of reports; train Comprehend custom model to recognize key oil & gas entities
- Apply model to auto tag documents with metadata like basins, lithology etc.

Knowledge Graph

- Work with geologists/engineers to identify key exploration concepts and relationships
- Model as nodes and edges in Neptune; visualize sample graph

Search & Query

- Index tagged datasets into Kendra engine to enable contextual search
- Build basic Lex chatbot leveraging cataloged schemas to take natural language queries

User Validation

- Provide user groups access to chatbot, knowledge graph and search interfaces
- Gather feedback on usability, accuracy and areas of improvement

Scale & Expand

- Expand custom Comprehend model with more vocabulary and document types
- Load more datasets into knowledge graph and Kendra index
- Enrich Lex bot with more supported query patterns

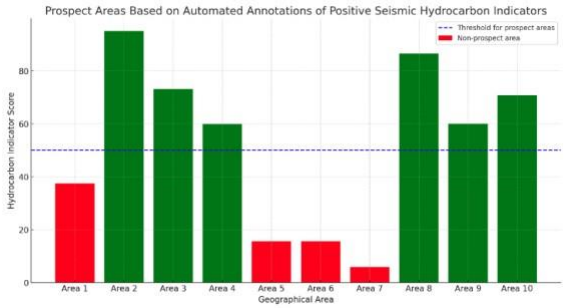
Success Metrics

- User satisfaction scores on self-service analytics experience
- Reduced time taken to find answers and insights

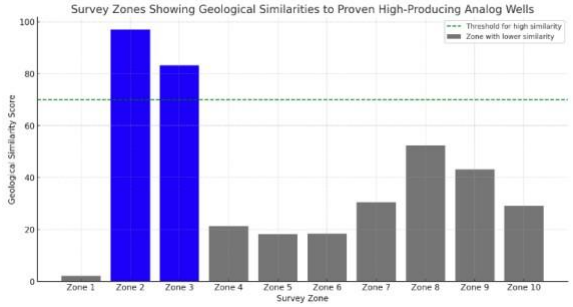
6. Uses

Here are potential business issues that could be analyzed from the ingested exploration data leveraging augmented analytics and NLP:

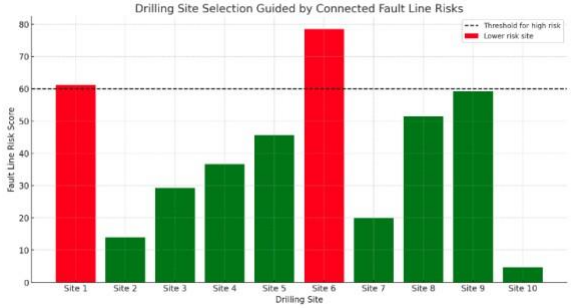
1. Identify prospect areas based on automated annotations of positive seismic hydrocarbon indicators



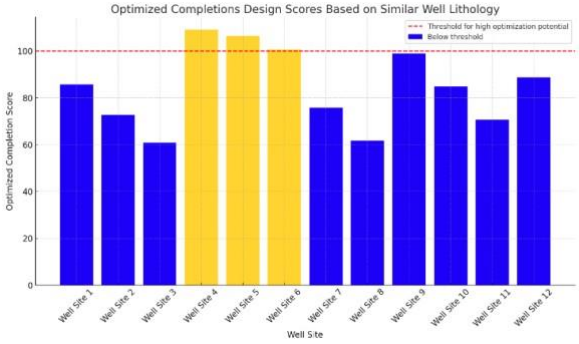
2. Highlight survey zones showing geological similarities to proven high-producing analog wells



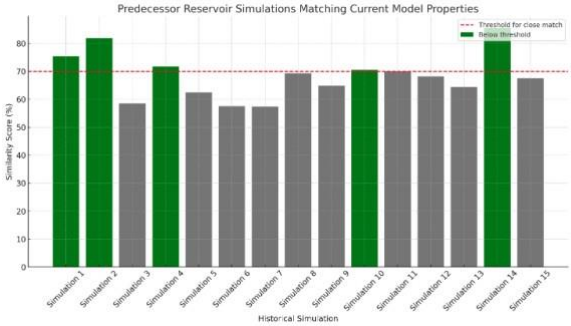
3. Guide drilling site selection by revealing previously unknown connected fault line risks



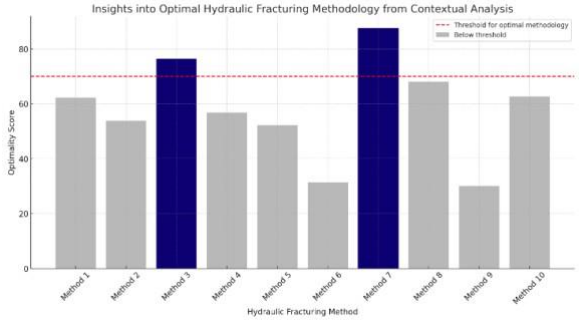
4. Recommend optimized completions design based on past techniques used in similar well lithology



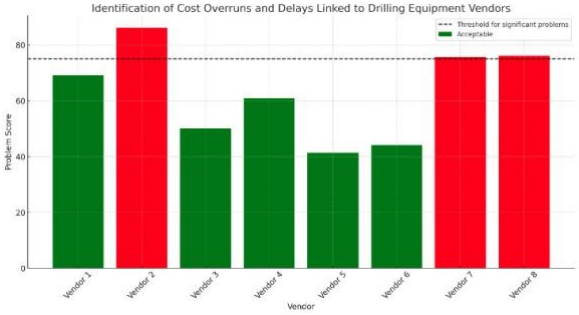
5. Retrieve examples of predecessor reservoir simulations matching current model properties



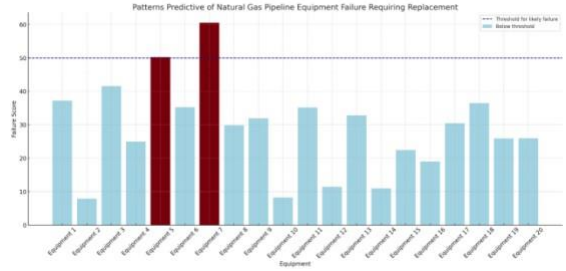
6. Gain insights into optimal hydraulic fracturing methodology from contextual analysis



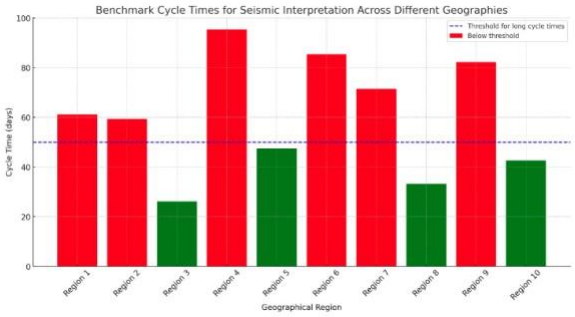
7. Discover cost overruns and delays linked to certain drilling equipment vendors



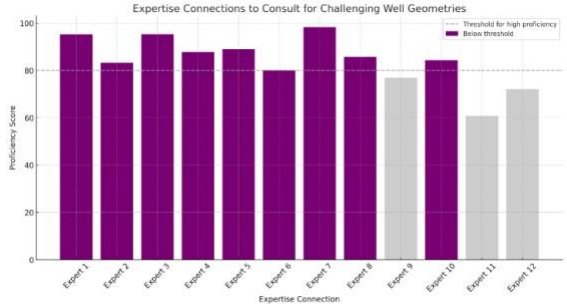
8. Pinpoint patterns predictive of natural gas pipeline equipment failure requiring replacement



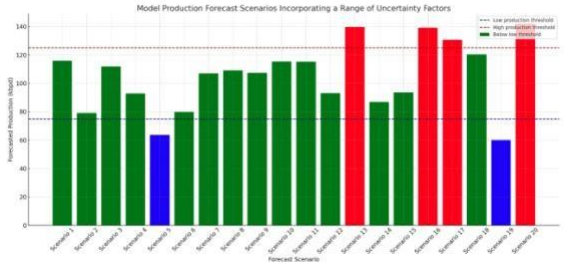
9. Benchmark cycle times for seismic interpretation across different geographies



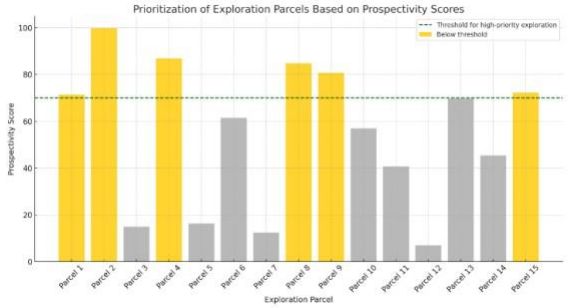
10. Uncover expertise connections to consult for challenging well geometries



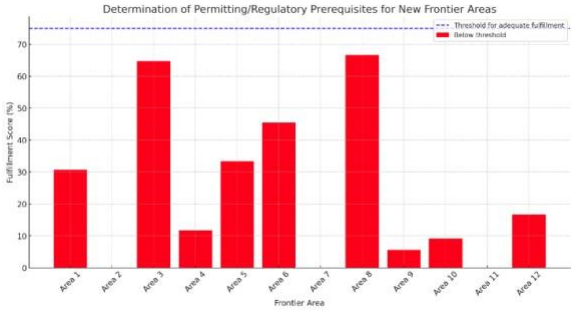
11. Model production forecast scenarios incorporating a range of uncertainty factors



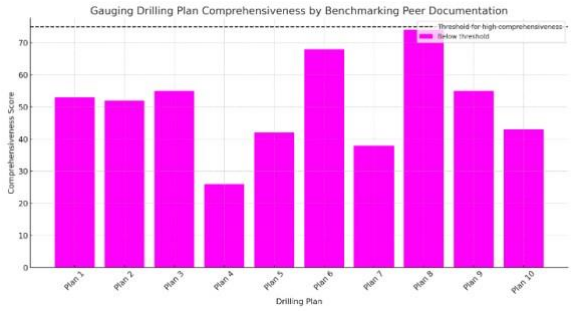
12. Prioritize exploration parcels based on automatically generated proceptivity scores



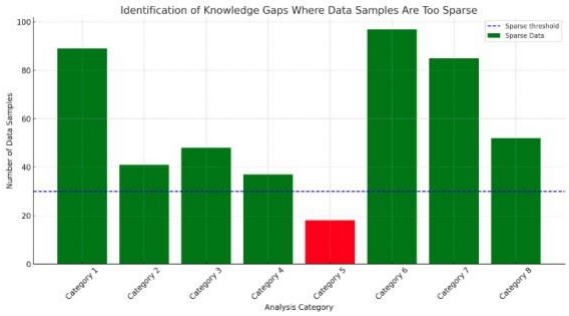
13. Determine permitting/regulatory prerequisites for exploring new frontier areas



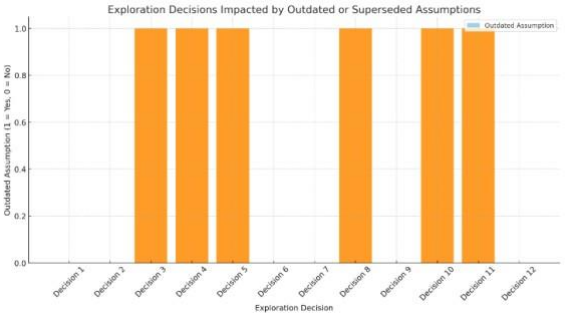
14. Gauge drilling plan comprehensiveness by benchmarking peer documentation



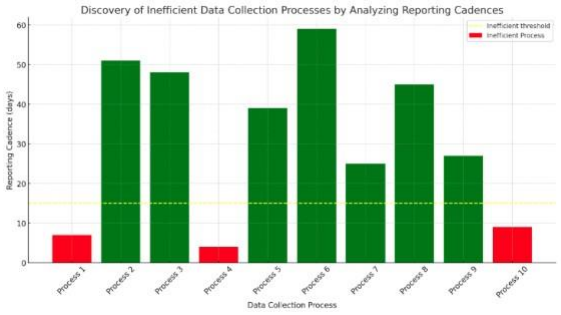
15. Identify knowledge gaps where data samples are too sparse to support sound analysis



16. Highlight exploration decisions impacted by outdated or superseded assumptions

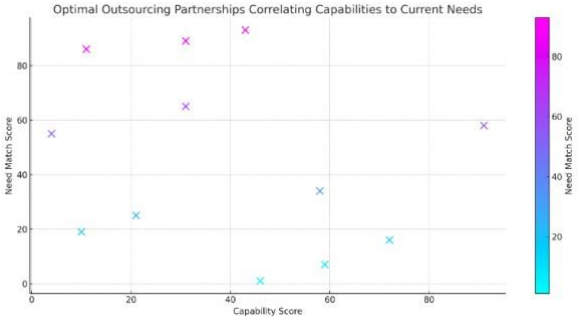


17. Discover inefficient data collection processes by analyzing reporting cadences

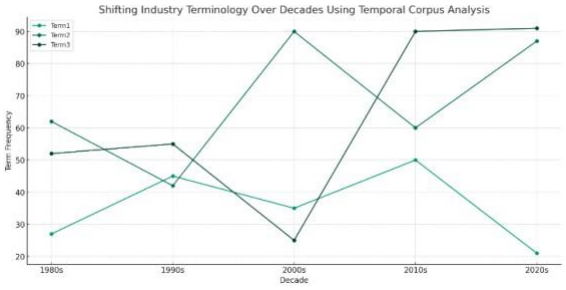


18. Extract learnings from costly failed projects applicable to new ventures

19. Recommend optimal outsourcing partnerships correlating capabilities to current needs



20. Detect shifting industry terminology over decades using temporal corpus analysis



7. Impact

Here are key business impacts that leveraging augmented analytics and NLP could drive for exploration insights:

1. Accelerate development planning and drilling decisions by enabling geologist's self-service access to vital engineering insights
2. Improve productivity of seismic analysis by automatically surfacing analogs tied to production outcomes
3. Boost collaboration and transfer of institutional knowledge across asset teams and geography silos

4. Reduce exploration program budget overruns through early discovery of higher-risk vendors
5. Enable competitive benchmarking by dynamically accessing peer performance datasets
6. Validate exploration investment assumptions and forecasts using automated reasoning engines
7. Create organization-wide transparency into expertise availability allowing better utilization
8. Drive higher staff productivity by eliminating delays waiting for analytics model translation
9. Uncover wasteful data collection habits allowing field optimization
10. Build a culture hungry for experimentation and evidence-based decisions through expanded analytics access

8. Extended Use Cases

Here are extended use cases leveraging augmented analytics and NLP across other industries:

1. Health:

Surface insights from patient history case files to improve clinical diagnosis and treatment planning accuracy.

2. Retail:

Extract shopper insights from customer service chat logs and feedback surveys to guide merchandising.

3. Travel:

Tap into contextual cues within traveler itineraries and journals to provide personalized recommendations.

4. Pharmacy:

Analyze pharmacological research paper corpora to accelerate drug discovery and repurposing opportunities.

5. Hospitality:

Harness food sustainability insights from restaurant reviews to shape venues' sourcing policies.

6. Supply Chain:

Interpret manufacturing sensor failure alerts using NLP to predict warranty issues.

7. Finance:

Democratize investment research by automatically linking market developments to stock implications.

8. E-Commerce:

Surface ecommerce site UX pain points from consumer reviews using sentiment analysis.

9. Shipping:

Identify vessel performance improvement areas by extracting insights from captain's logs.

10. CRM:

Boost customer intelligence by dynamically profiling personas using interaction transcript analysis.

9. Conclusions

The sheer amount of complicated, raw engineering information produced throughout the stages of oil and gas discovery has surpassed the capabilities of manual analysis by humans, leading to bottlenecks in accessing insights.

The need for specialized expertise to decipher signals from seismic data or reservoir simulations adds to the constraints on efficiency.

Proposed in this study is an enhanced analytics structure aimed at tackling these issues through the integration of machine learning pathways for automatic tagging of metadata, knowledge graphs for linking entities, and user- friendly conversational interfaces for easy queries.

A cloud-service model showed the potential of custom- built language processors to precisely interpret technical documents related to oil and gas, bringing crucial insights to the forefront. The extraction of entities helped fill interactive databases, facilitating autonomous exploration of links. Additionally, chatbots allowed users to receive responses by posing questions in everyday business terminology.

Further comparisons highlighted enhancements in the data collection process.

The document outlines how adopting augmented analytics innovations can break free the locked value in unstructured engineering information. The strategies for making data insights more widely accessible are relevant in various industrial settings that wrestle with escalating data amounts and limited expertise, speeding up decision- making.

Further advancements are planned to improve the generation of insights autonomously through the application of reinforcement learning on user interaction patterns. Merging with operational data streams could enhance relevance. Expanding these platforms to incorporate collective knowledge has the potential to boost the adaptability of organizations in the dynamic global energy sector.

References

[1]. Georgiou, K., Mittas, N., Mamalikidis, I., Mitropoulos, A. C., & Angelis, L. (2021). Analyzing the roles and competence demand for digitalization in the oil and gas 4.0 era. IEEE Access, 9, 151306–151326. https://doi.org/10.1109/access.2021.3124909

[2]. Suiçmez, V. S. (2019). What does the data revolution offer the oil industry? Journal of Petroleum Technology, 71(03), 33. https://doi.org/10.2118/0319-0033-jpt

[3]. Alnuaim, S. (2018). Energy, Environment, and Social Development: SPE's new Strategic Plan - emphasizing pride in what we do. Journal of Petroleum Technology, 70(11), 10–11. https://doi.org/10.2118/1118-0010-jpt

[4]. Alnuaim, S. (2018). Energy, Environment, and Social Development: SPE's new Strategic Plan - emphasizing pride in what we do. Journal of Petroleum Technology, 70(11), 10–11. https://doi.org/10.2118/1118-0010-jpt

[5]. Da Silva Magalhães Gomes, D., Cordeiro, F. C., Consoli, B. S., Santos, N. L., Moreira, V. P., Vieira, R., Moraes, S. M. W., & Evsukoff, A. G. (2021). Portuguese word embeddings for the oil and gas industry: Development and evaluation. Computers in Industry, 124, 103347. https://doi.org/10.1016/j.compind.2020.103347

[6]. Georgiou, K., Mittas, N., Mamalikidis, I., Mitropoulos, A. C., & Angelis, L. (2021). Analyzing the roles and competence demand for digitalization in the oil and gas 4.0 era. IEEE Access, 9, 151306–151326. https://doi.org/10.1109/access.2021.3124909

[7]. Zhou, X., Gong, R., Fugeng, S., & Wang, Z. (2020). PetRoKG: Construction and application of Knowledge Graph in upstream area of PetroChina. Journal of Computer Science and Technology, 35(2), 368–378. https://doi.org/10.1007/s11390-020-9966-7

[8]. Staar, P., Dolfi, M., & Auer, C. (2020). Corpus processing service: A Knowledge Graph platform to perform deep data exploration on corpora. Applied AI Letters, 1(2). https://doi.org/10.1002/ail2.20

[9]. Georgiou, K., Mittas, N., Mamalikidis, I., Mitropoulos, A. C., & Angelis, L. (2021). Analyzing the roles and competence demand for digitalization in the oil and gas 4.0 era. IEEE Access, 9, 151306–151326. https://doi.org/10.1109/access.2021.3124909

[10]. Chen, X., Xie, J., Gao, J., Wang, R., & Jiang, J. (2021). Dynamic knowledge graph-based construction of quality infrastructure system for Non-API oil Country tubular goods. Journal of Physics: Conference Series, 2101(1), 012041. https://doi.org/10.1088/1742-6596/2101/1/012041

[11]. Sagdatullin, A. (2021). Investigation of the Possibility of Building a Neural Fuzzy Logic Controller with Discrete Terms for Controlling and Automating Oil and Gas Engineering Facilities. Интеллектуальные Системы В Производстве, 19(3), 105–110. https://doi.org/10.22213/2410-9304-2021-3-105-110

[12]. Wilson, S. (2021). Technology Focus: Gas Production (August 2021). Journal of Petroleum Technology, 73(08), 62. https://doi.org/10.2118/0821-0062-jpt

[13]. Zhou, B., Svetashova, Y., De Gusmão, A. L., Soylu, A., Cheng, G., Mikut, R., Waaler, A., & Kharlamov, E. (2021). SemML: Facilitating development of ML models for condition monitoring with semantics. Journal of Web Semantics, 71, 100664. https://doi.org/10.1016/j.websem.2021.100664

[14]. Zhou, F., He, Y., Ma, P., & Mahto, R. V. (2020). Knowledge management practice of medical cloud logistics industry: transportation resource semantic discovery based on ontology modelling. Journal of Intellectual Capital, 22(2), 360–383. https://doi.org/10.1108/jic-03-2020-0072

[15]. Shi, W., Tang, D., & Zou, P. (2021). Research on cloud enterprise resource integration and scheduling technology based on mixed set programming. Tehnicki Vjesnik-technical Gazette, 28(6). https://doi.org/10.17559/tv-20210718091658

[16]. Dwivedi, A. K., & Satapathy, S. M. (2021). Ontology-Based modelling of IoT design patterns. Journal of Information & Knowledge Management, 20(Supp01), 2140003. https://doi.org/10.1142/s0219649221400037

[17]. Preisig, H. A. (2021). Ontology-Based Process modelling-with examples of physical topologies. Processes, 9(4), 592. https://doi.org/10.3390/pr9040592

[18]. Tian, Y., Gao, J., Liu, N., & Chen, D. (2021). Construction of optimal basic Wavelet via AIDNN and its application in seismic data analysis. IEEE Geoscience and Remote Sensing Letters, 18(7), 1144–1148. https://doi.org/10.1109/lgrs.2020.2997339

[19]. Cianetti, S., Bruni, R., Gaviano, S., Keir, D., Piccinini, D., Saccorotti, G., & Giunchi, C. (2021). Comparison of deep learning techniques for the investigation of a seismic sequence: an application to the 2019, MW 4.5 Mugello (Italy) earthquake. Journal of Geophysical Research: Solid Earth, 126(12). https://doi.org/10.1029/2021jb023405

[20]. Konstantaras, A., Petrakis, N. S., Frantzeskakis, T., Markoulakis, E., Kabassi, K., Vardiambasis, I. O., Kapetanakis, T. N., Moshou, A., & Maravelakis, E. (2021). Deep learning neural network seismic big-data analysis of earthquake correlations in distinct seismic regions. International Journal of Advanced Technology and Engineering Exploration, 8(84). https://doi.org/10.19101/ijatee.2021.874641

[21]. Ali, K. K., Wanas, A., & Mahdi, M. E. (2021). Application of velocity analysis picking for 2D seismic data processing in West An-Najaf are. Iraqi Journal of Science, 555–564. https://doi.org/10.24996/ijs.2021.62.2.21

[22]. Ali, K. K., Wanas, A., & Mahdi, M. E. (2021). Application of velocity analysis picking for 2D seismic data processing in West An-Najaf are. Iraqi Journal of Science, 555–564. https://doi.org/10.24996/ijs.2021.62.2.21